PG program in Big Data & Analytics

Data Analytics Applications Project Report

Predicting earning potential on Adult Dataset

&

Hyper parameter tuning for XGBOOST

Submitted by: Dattatray Shinde

Mentor: Prof. Soumendra

Submission date 15/03/2017

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PG certification in Big Data & Analytics from S P Jain School of Golbal Mangement is entirely my own work except where otherwise stated, and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfillment of the requirements of that stated above.

Author: Dattatray Shinde Dated: 15/03/2017

Abstract

This report implements the CRISP-DM methodology when applying classification models to the problem of identifying individuals whose salary exceeds a specified value based on demographic information such as age, level of education and current employment type. The process involved in the exploration, preparation, modelling and evaluation of the datasets are described. Topics such as the application of statistical analysis to suggest attribute usefulness, feature reduction, outlier detection, missing value management, data bias and data transformation is discussed. The process of relative performance analysis of the proposed classifiers is reviewed. The support of a business objective which will use the predictive capabilities of the proposed models to target customers is reviewed including the use of lift analysis to indicate the likely level of return on investment and overall profitability.

Table of Contents

Introduction

This project will investigate the data mining of demographic data in order to create one or more classification models which are capable of accurately identifying individuals whose salary exceeds a specified value. The data used in this project were sourced from the University of California Irvine data repository and are referred to as the Adult dataset and contain information on individuals such as age, level of education and current employment type.
The classification model will be used to select candidates for a new service offered by the sponsor of the project targeting individuals with salaries exceeding fifty thousand US dollars. A description of the predictive significance of each attribute, interesting or useful patterns which were found and any transformations applied to the data must be provided with all proposed models.

This report will describe the work carried out during the iterative process of data preparation, modelling and evaluation including data formatting, consistency or other quality issues, opportunities for useless instance or attribute removal and the approaches taken to solving issues with instances affected by noise, outliers or missing values.

This project will implement the CRISP-DM methodology where a comprehensive review of the customer's requirements supports the creation of a business objective outlining items such as the expected level of model performance and return on investment. This will be used to create a data mining objective which will guide the subsequent work in the data understanding, preparation and modelling steps and the final evaluation and selection (after revisiting earlier steps if necessary) of a classification model or models.

1. Business Understanding

Business objective

A project team has been created to support the marketing of a new service targeted at potential customers with medium to high level salaries. There is an initial project setup cost of eighteen thousand dollars, a cost of one hundred and twenty five dollars for each offer made and a return of five hundred dollars for each accepted offer:

Setup cost:	$18,000
Cost per offer:	$125
Return per accepted offer:	$500

The current marketing strategy, which involves a high degree of investment per offer (driving the relatively high offer cost), has achieved a high acceptance rate in the past of seventy five percent by individuals whose salaries exceed fifty thousand US dollars. The goal of this project is therefore to create a model (or models) which can accurately identify individuals whose annual salary exceeds this amount. Any proposed model must be capable of significantly outperforming the existing candidate selection model which is currently providing an average return on investment of seventy to ninety.

Data mining objective

The data mining objective is to create a classification model which can predict individuals whose salary exceeds fifty thousand US dollars by mining anonymised census data containing demographic information such as age, gender, and education level and employment type. The original salary attribute in the census data has been anonymised to a binomial value indicating if a salary exceeds fifty thousand US dollars.

For each proposed model an expected return on investment (and associated profit margins) must be provided. A clear description of all data transformations which were carried out must be provided and any useful insights into the data such as significant attributes or mining issues within the data should be included.

2. EDA (Exploratory Data Analysis)

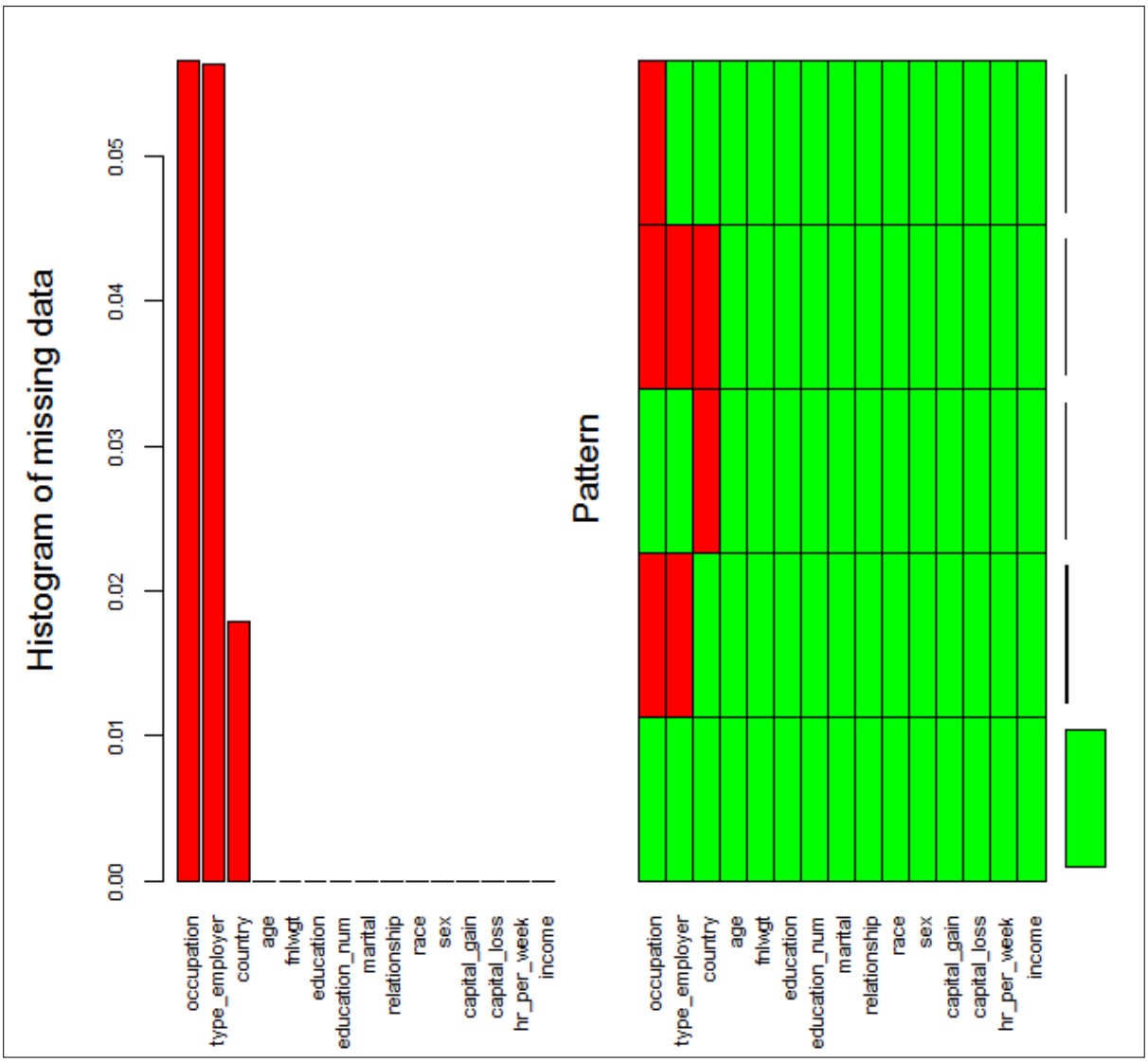
1. Variable Identification

The dataset used in this project has 49000 records and a binomial label indicating a salary of less or greater than $50,000 dollars, which for brevity, will be referred to as <50K or >50K in this report. 76 % of the records in the dataset have a class label of <50K. The data has been divided into a training dataset containing 32000 records and a test dataset containing 16000 records.

There are 14 attributes consisting of 7 polynomials, 1 binomial and 6 continuous attributes. The nominal employment class attribute describes the type of employer such as self employed or federal and occupation describes the employment type such as farming or managerial. The education attribute contains the highest level of education attained such as high school graduate or doctorate. The relationship attribute has categories such as unmarried or husband and the marital status attribute has categories such as married or separated. The final nominal attributes are country of residence, gender and race. The continuous attributes are age, hours worked per week, education number (which is a numerical representation of the nominal education attribute), capital gain and loss and a survey weight attribute which a demographic score is assigned to an individual based on information such as area of residence and type of employment.

2. Missing Data Analysis (R Package Mice)

As we can see from following bar chart that more than 5% of data is missing for occupation and employer type. Less than 2% data is also missing for country which we will impute in later stages if required.



Univariate & Bivariate Data Analysis

While doing EDA, removed all missing values which we can impute in modelling if after imputation we get the better results.

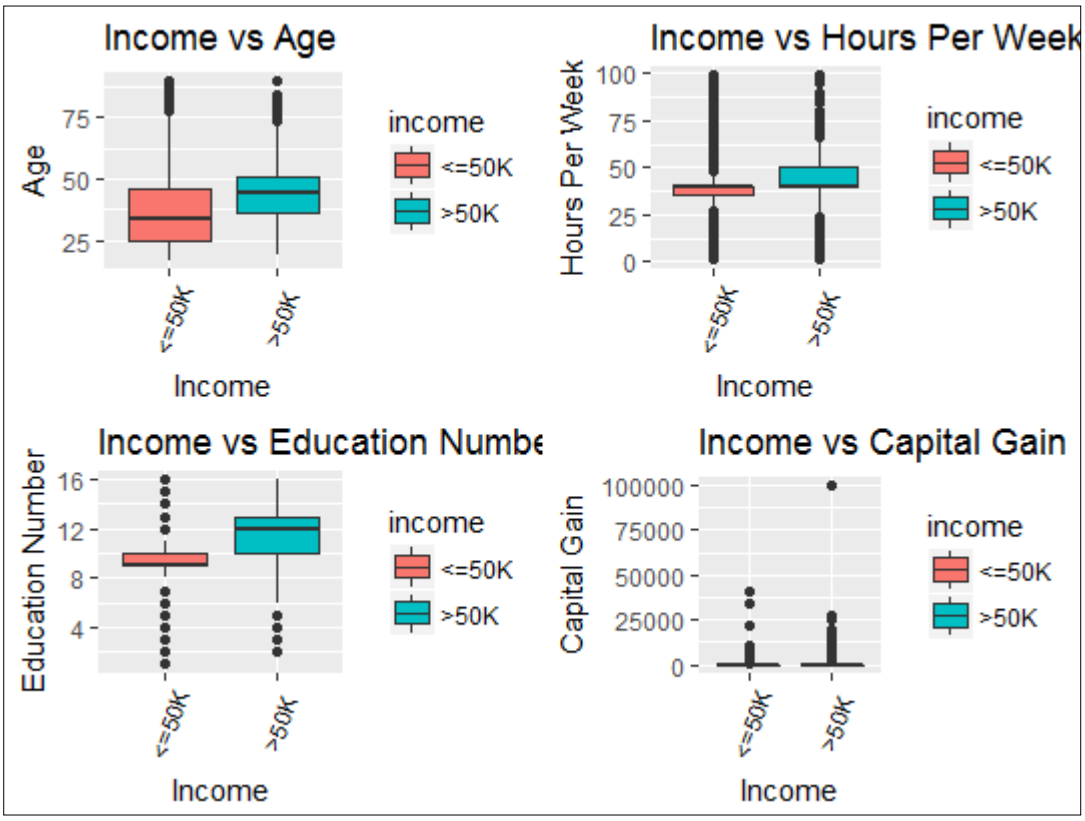
1: Descriptive Statistics

	Attribute	Values						Missing
Polynomials	Employment Class	Private (68%), Self employed 1 (8%), Local Gov(6%), State Gov(4%), Unknown (5%), Self employed 2 (3%), Federal Gov(3%), No Pay(1.5%), Never Worked (0.5%)						1836
	Education Level	High School (32%), Some college (22%), Bachelors (16%), Masters (5%), Vocational (4%), 11th (4%), Assoc Academic (3%), 10th (3%), 7-8th (2%), Professional School (2%), 9th (2%), 12th (2%), Doctorate (1%), 5-6th (1%), 1-4th (1%), Preschool (1%)						0
	Relationship	Husband (41%), Not-in-family (26%), Own child (16%), Unmarried (11%), Wife (4%), Other relative (2%)						0
	Race	White (85%), Black (10%), Asian / Pacific Islander (3%), American Indian / Eskimo (1%), Other (1%)						0
	Marital Status	Married-civ-spouse (46%), Never-married (33%), Divorced (14%) Separated (3%), Widowed (2%), Married-AF-spouse (1%), Married-spouse-absent (1%)						0
	Occupation	15 categories						1843
	Country	42 categories: USA (90%)						583
Binomials	Salary [Label]	<=$50K (76%), >$50K (24%)						0
	Gender	Male (67%), Female (33%)						0
Real		Mean	Median	Std Dev	Skewness	Kurtosis	Range	
	Age	38.58	37	13.64	0.56	2.83	17 - 90	0
	Hours worked per week	40.44	40	12.35	0.23	5.92	1 - 99	0
	Education Number	10.08	10	2.57	-0.31	3.62	1 - 16	0
	Captial Gain	1078	0	7385	11.95	157.77	0 - 99999	0
	Capital Loss	87.3	0	403	4.59	23.37	0 - 4356	0
	Survey Weight	189778	178356	105550	1.45	9.22	12285 - 1484705	0

The standard deviations in the data above indicates that there is a significant quantity of values in all attributes, particularly in the case of the survey weight attribute and the high kurtosis in the capital gain attribute indicates a long tail. From the boxplots it can be seen that the range of attribute values for age, education number and hrs_per_week (worked) is slightly higher in >50K class instances (Point #2) indicating that these attributes may have predictive significance.

2: Bi-variate analysis

It will be better to have Box plots, Scatter Plots , Histograms , heat map so that we can easily identify relationship between features as well as it will help us to identify outliers and significance of features in classification



The above box plots will helps us to identify the outliers as well as the features which will be significant in the classification.

The numeric attributes appear to contain a significant quantity of unique values and in the case of the survey weight attribute there were twenty one thousand unique values out of thirty one thousand instances, which may suggest that this attribute may not be significantly predictive. This was confirmed when a regression model was applied to the dataset where fnlwgt had a t-Statistic of zero and a p-Value of one (Figure 2). The other attributes were found to be reasonably significant with the exception of the country attribute (which contains a value of 'United-States' in ninety percent of instances) and the race attribute (which contains a value of 'White' in eighty five percent of instances), and may be candidates for removal during data preparation, and the marital status attribute which is explored further later in this report.

although the capital gain and loss attributes have significant quantities of unique values (Table 2), the majority of the instances have a zero value with capital gain having ninety six percent zero values in the <50K class and seventy nine percent in the >50K class and capital loss having ninety eight percent in the <50K class and ninety percent in the >50K class. This may indicate that these attributes also may not be very predictive.

The numeric education number and nominal education level attributes were found to be fully correlated and therefore one of these attributes may be a good candidate for removal during modelling (Table 3). In general the other attributes were found to be weakly correlated (Figure 3). When the education number attribute was plotted for the class labels it was found that the lower values tend to predominate in the <50K class and higher levels in the >50K class which may indicate some predictive capability (Figure 4).

Univariate data analysis

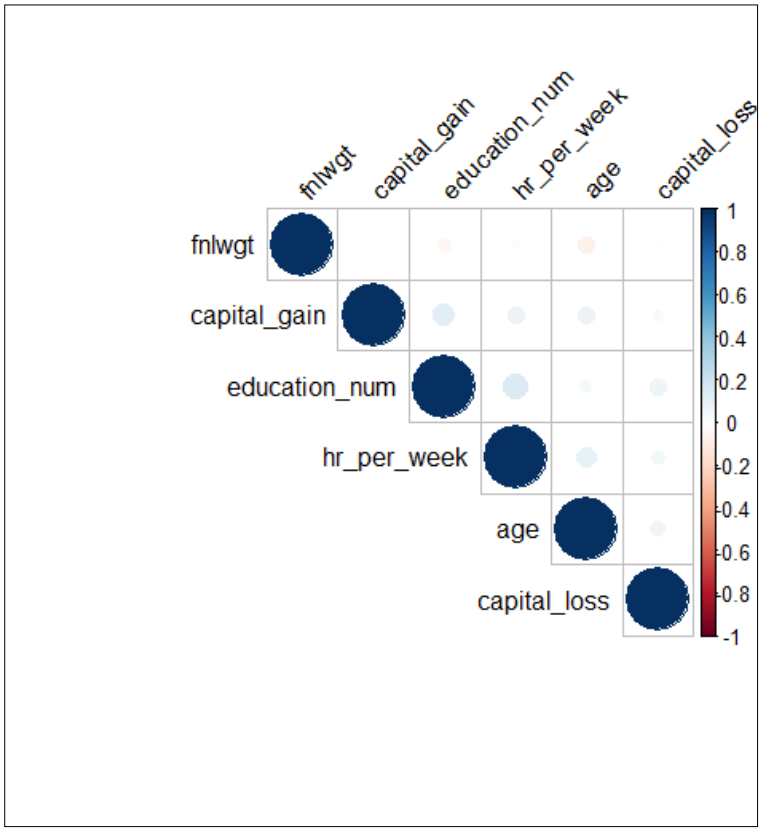
Mapping between education number and education level attributes

The numeric education number and nominal education level attributes were found to be fully correlated and therefore one of these attributes may be a good candidate for removal During modelling (Table 3). In general the other attributes were found to be weakly Correlated (Figure 3). When the education number attribute was plotted for the class labels it was found that the lower values tend to predominate in the <50K class and higher levels in the >50K class which may indicate some predictive capability.

Educ Num	1	2	3	4	5	6	7	8
Education	Preshool	1st-4th	5th-6th	7th-8th	9th	10th	11th	12th
Educ num	9	10	11	12	13	14	15	16
Education	HS-grad	Some-college	Assoc-voc	Assoc-acdm	Bachelors	Master	Prof-school	Doctorate

Correlation matrix for the training dataset (R Package corrplot)

With correlation matrix we tried to find the correlation between the variables. But the correlation matrix is showing no correlation as such visibly. So, no need to drop any variables. IN modelling step we will take vif(variance inflation factor) to check multi-co linearity again.

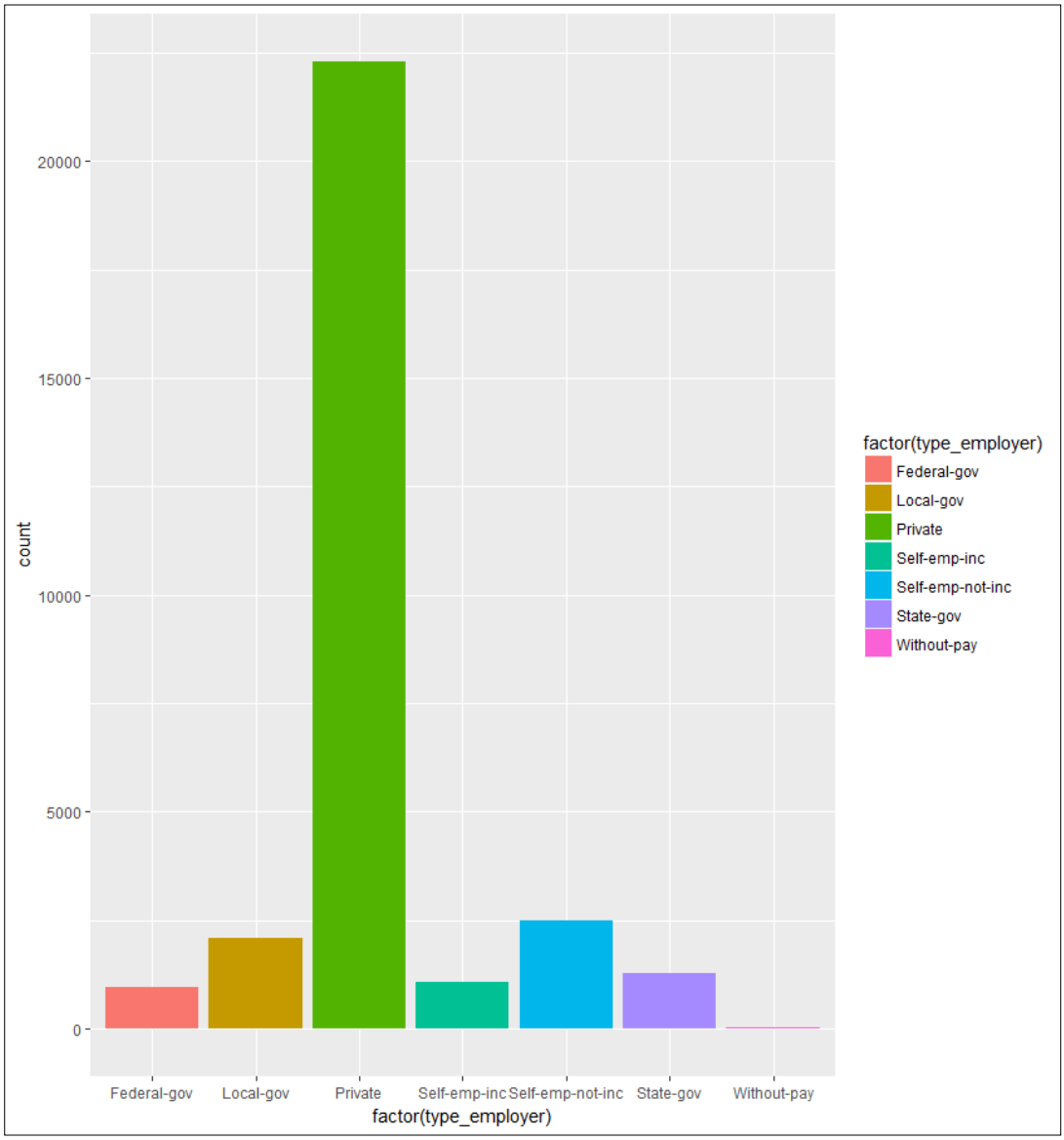


Employer Type Transformations

Looking at the bar chart below, we can easily figure out that “private” employer type is dominating here. Also given "Never worked" and "Without-Pay" are both very small groups, and they are Likely very similar, we can combine them to form a "Not Working" Category. In a similar vein, we can combine government employee categories, and self-employed categories. This allows us to reduce the number of categories significantly.

So the final factors for category employer type will be -

- Federal-Govt
- Other-Govt
- Private
- Self-Employed
- Not-Working

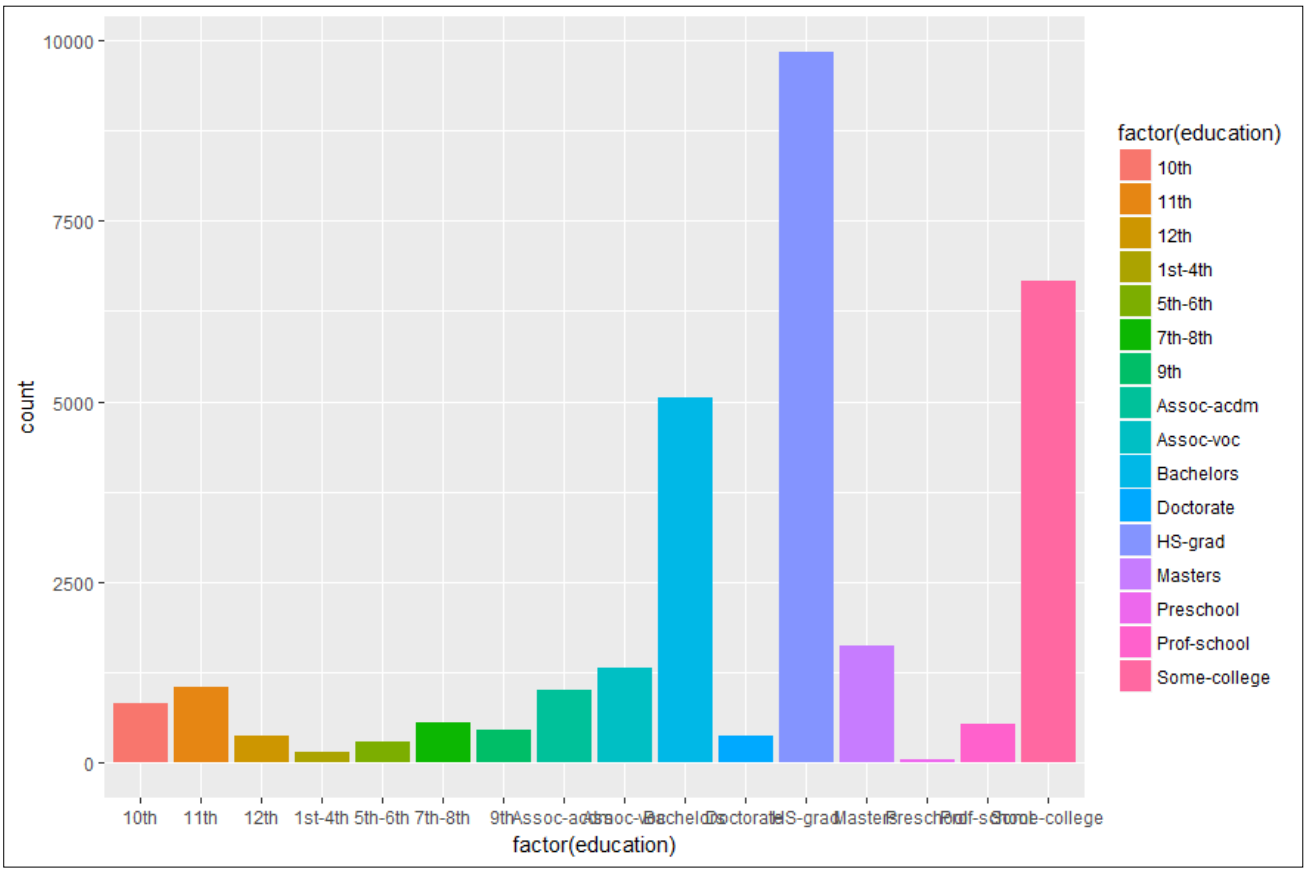


Education Level transformation

Block all the dropouts together. Block high school graduates and those that attended some college without receiving a degree as another group. Those college graduates who receive an associates are blocked together regardless of type of associates. Those who graduated college with a Bachelors, and those who went on to graduate school without receiving a degree are blocked as another group. Most everything thereafter is separated into its own group.

Final categories for education level feature will be as follows -

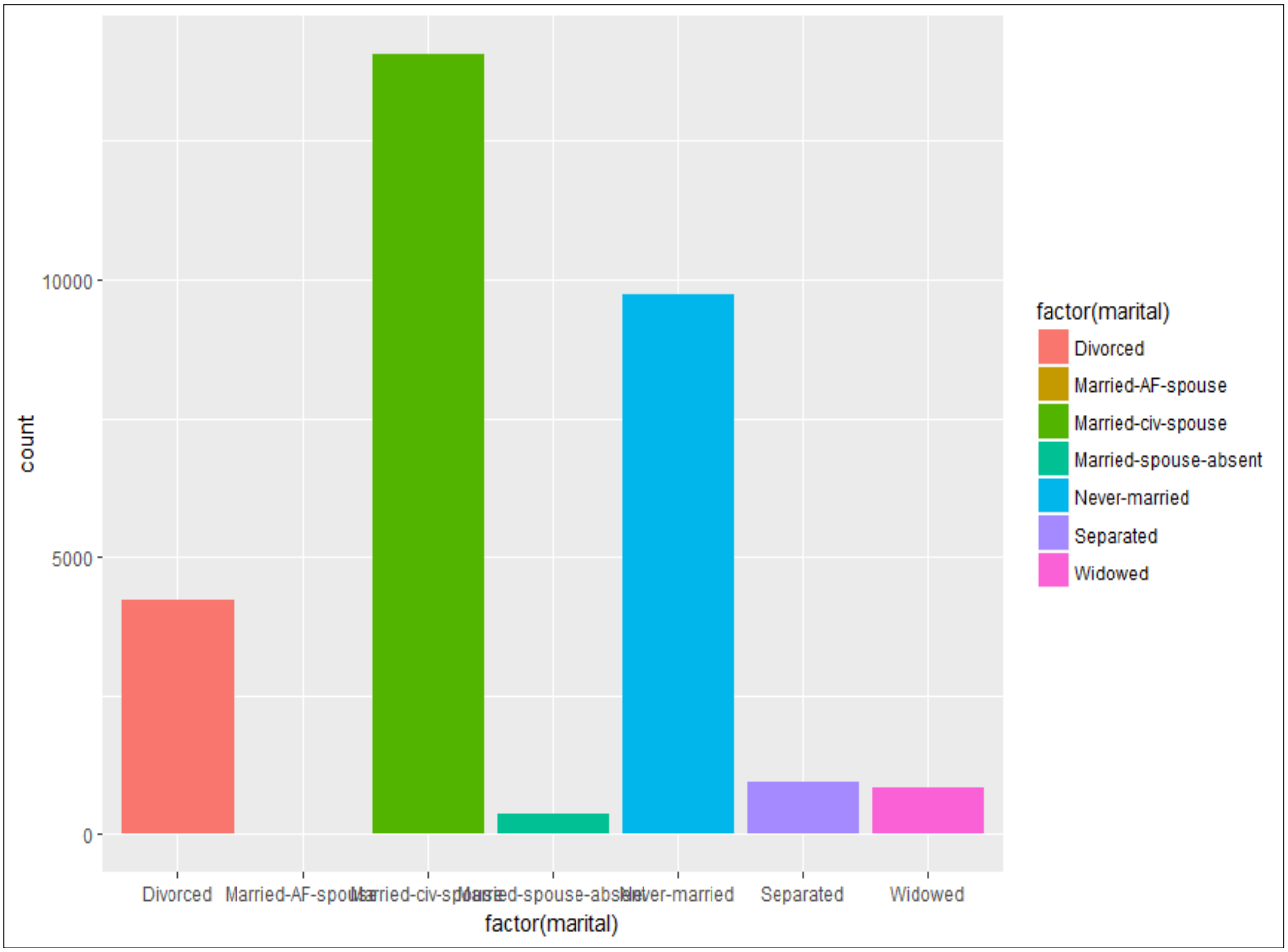
- 10^{th},11th,12th,1st-4th,5th-6th ,7th-8th,9th, Preschool added to new Dropout category
- Assoc-acdm, Assoc-voc added to Associates
- Bachelors to Bachelors
- Doctorate to Doctorate
- HS-Grad, Some-college added to HS-Graduate
- Masters to Masters



Marital Status transformations

Created new category Not-Married and done following changes –

- Married-AF-spouse added to Married
- Married-civ-spouse added to Married
- Married-spouse-absent added to Not-Married
- Separated added to Not-Married
- Divorced added to Not-Married

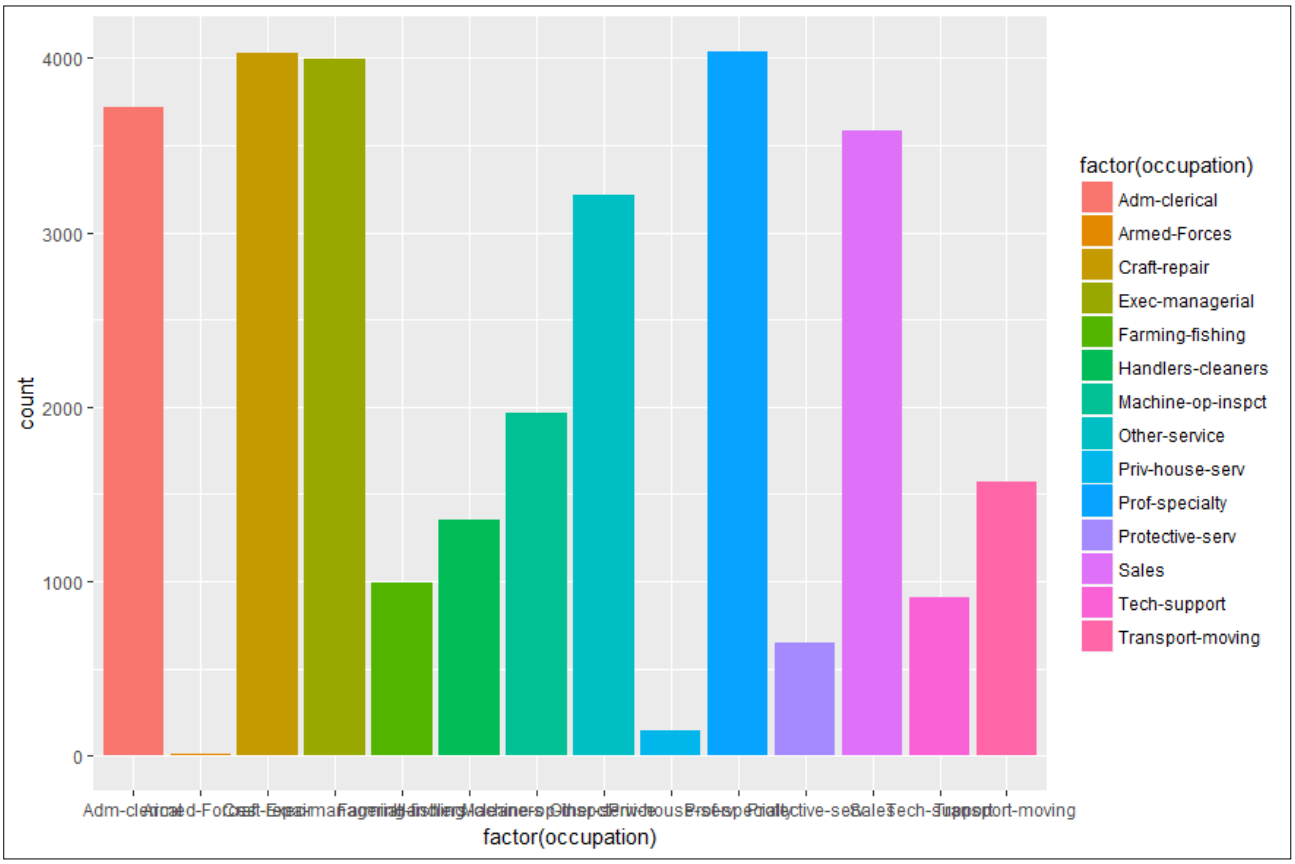


Occupation

The occupation attribute was found to have some very infrequently occurring values such as Armed-Forces and Priv-house-serv, some values such as Cleaners and other which are more highly correlated with the < 50K class and other values such as Exec-managerial and Prof-specialty being more highly correlated with the < 50K class. This correlation between occupation categories and the class label (in particular the >50K label) was also observed when a GLM was applied to the training dataset which suggests that this attribute may have a relatively high level of predictive capability.

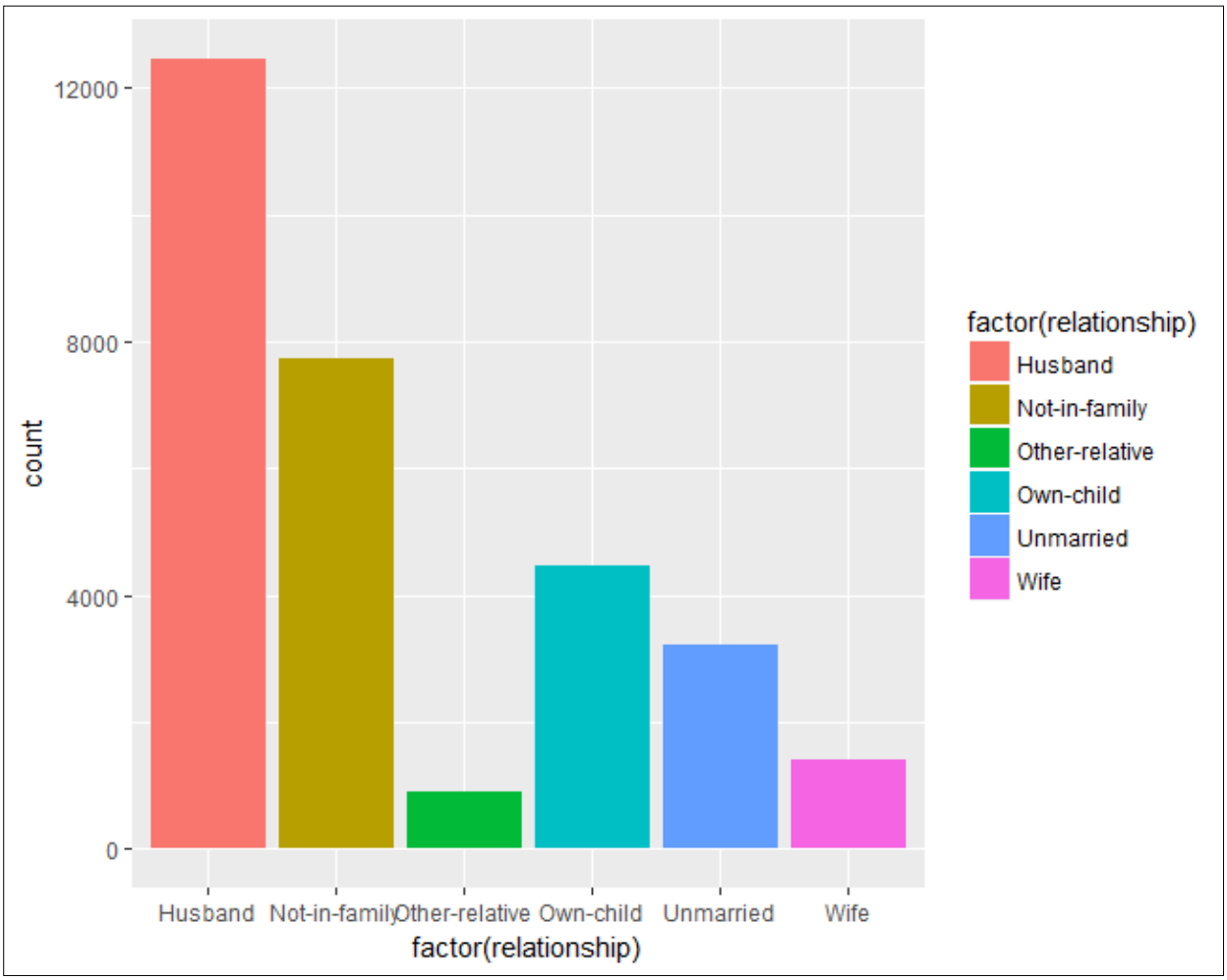
Given "Never worked" and "Without-Pay" are both very small groups, and they are likely very similar, we can combine them to form a "Not Working" Category.
In a similar vein, we can combine government employee categories, and self-employed categories. This allows us to reduce the number of categories significantly.
The changes done to the categories are as follows -

- Federal-gov to Federal-Govt
- Local-gov to Other-Govt
- State-gov to Other-Govt
- Private to Private
- Self-emp-inc to Self-Employed
- Self-emp-not-inc to Self-Employed
- Without-pay to Not-Working
- Never-worked to Not-Working



Relationship Transformations

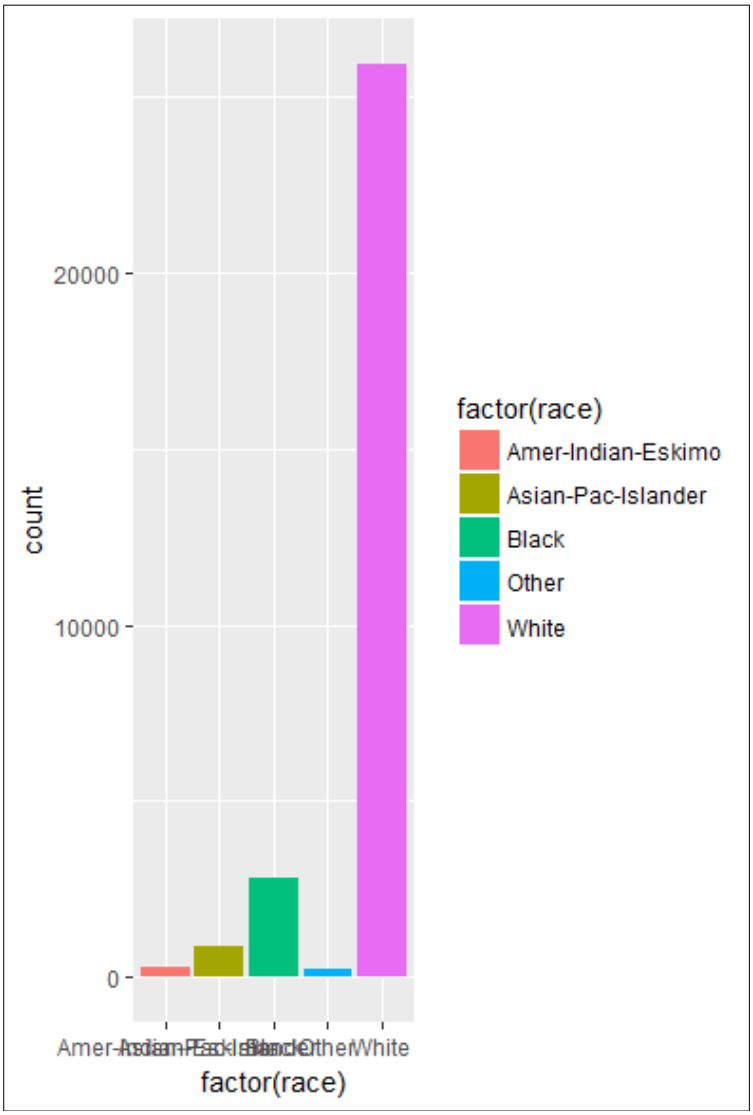
Nothing has been changed as far as relationship has been concerned.



Race category analysis

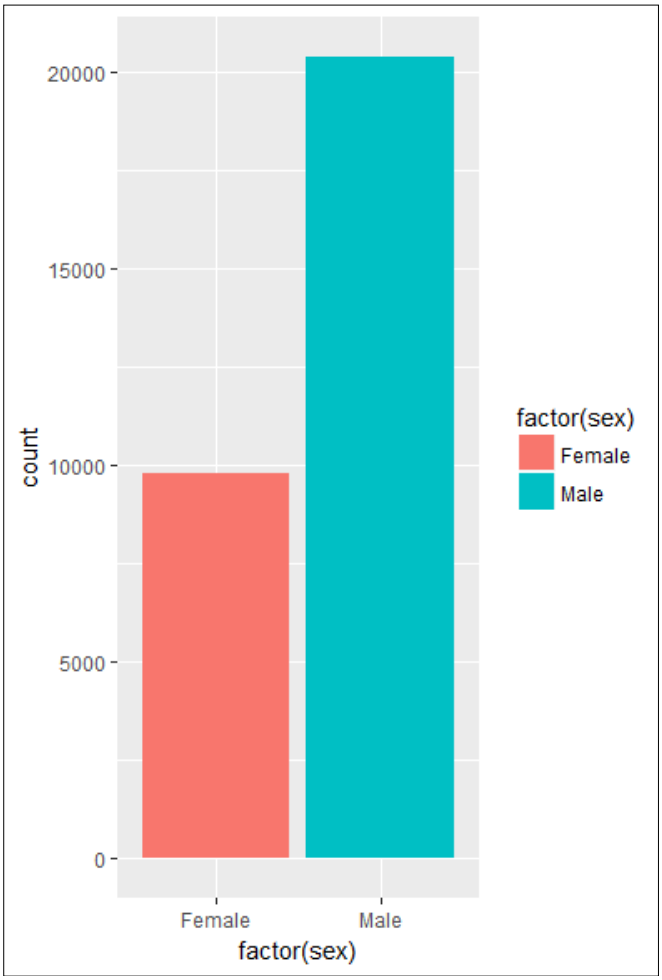
We can get an idea that most of the race value is “white” and it is dominating. We have just changed few factor names as follows for race category –

- Amer-Indian-Eskimo as Amer-Indian
- Asian-Pac-Islander as Asian



Sex category transformation

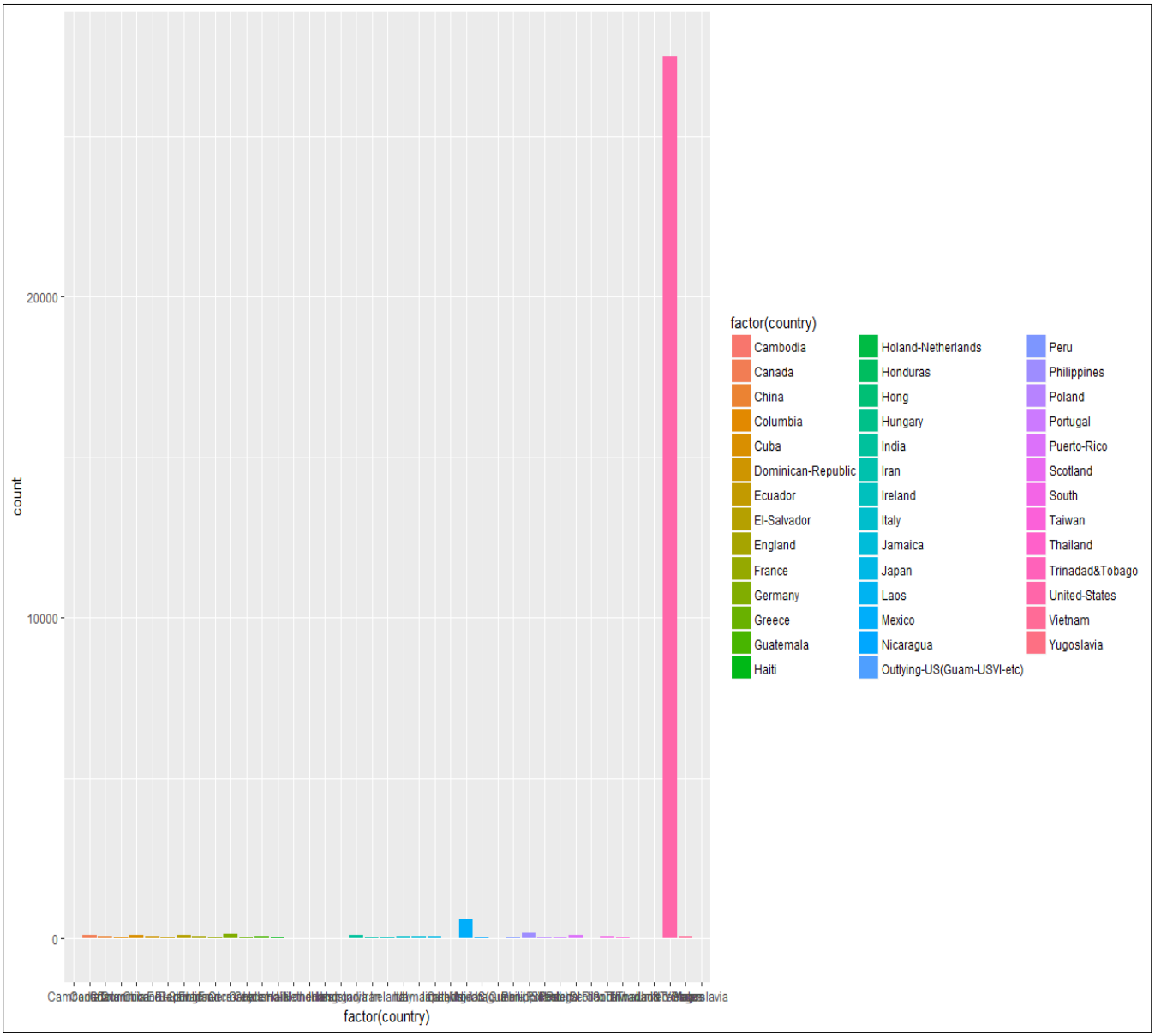
There is nothing to transform or normalize as far as sex data is concern. A slight bias was detected in the dataset where instances with a 'female' gender value have lower range of age values than instances with a 'male' gender value. This may skew the predictive capability of the gender attribute to some degree as age values within >50K class instances tend to be higher than <50K instances. There is also a slight imbalance in gender with sixty seven percent of instances having a male value.



Transformation in Country feature

The variable country presents a small problem. Obviously the United States represents the vast majority of observations, but some of the groups have such small numbers that their contributions might not be significant. A way around this would be to bin the countries by using a combination of geographical location, political organization, and economic zones.

- Euro_1 is countries within the Eurozone that are considered more affluent, and therefore people from there are probably going to be more affluent.
- Euro_2 - includes countries within the Eurozone that are considered less affluent. These included countries that are financially troubled like Spain and Portugal, but also the Slavic countries and those formerly influenced by the USSR like Poland.
- British-Commonwealth - Formerly British holdings that are still closely economically aligned with Britain are included under the British-Commonwealth.



Comparison of occupation attributes values within the salary classes. A positive percentage change indicates an increased proportion in the >50K class.

Correlation of occupation categories with >50K class from decision tree on training dataset.

In order to test the observations outlined above a rule induction classifier was applied to the training dataset which achieved an accuracy of 82.85%. The runtime was quite long at three hours and thirty minutes so this classifier would not be the optimal choice on this dataset unless it could be proven to significantly outperform other classifiers. The generated rules were very useful however in confirming some of the observations already noted such as the correlation between the occupation attribute and the >50K salary class (Table 5).

Oversampling with SMOTE

I was thinking to have oversampling as class >=50K is 75 % in data and remaining in <50K.So, by random guess also one can have 0.75 accuracy of predicting classes. But I have not chosen to oversample as basic models like Naive Bayes also giving probability more than 85 %.

Verification of data quality

In general it was found that the incidence of extreme outliers in the dataset was low with the exception of the hours_per_week (worked) attribute which has a significant percentage of outliers in both tails. The age and fnlwgt (survey weight) attributes had outliers in less than three percent of instances (Table 2). Rapidminer's Detect Outlier (Distances) operator (with a k value of six based on kNN classifier modelling) detected outliers primarily in the fnlwgt (survey weight) and capital gain attributes (Figure 7).

There are 3 attributes with missing values with an incidence rate of two percent for the country attribute and four percent for employment class and occupation (Table 1). It may be possible to impute the missing country values as this has a 'United-States' value ninety percent of the time and for employment class with 'Private' occurring in seventy three percent of instances. Imputing values for the occupation attribute may be more challenging as this attribute's values are more evenly distributed. As the overall incidence rate of missing values is quite low at less than five percent, it may be possible to remove instances with missing values without the need for value imputation or replacement.

The dataset was also checked for inconsistencies and conflicts such as male gender with a relationship value of wife or unmarried marital status with a relationship value of husband/wife etc, but very few instances of this type of issue were found indicating that the data is of a reasonable quality. With thirty two thousand available instances (or thirty thousand if all missing value instances were removed) and low levels of data duplication (only twenty four duplicate instances exist in the training dataset) there should be a sufficient quantity of clean data available to learn a classification model. There is some class imbalance within the dataset as 76 % instances are in the <50K class, but the 8000 instances in the >50K class should be sufficient to detect the patterns within this class or otherwise boosting can be applied.

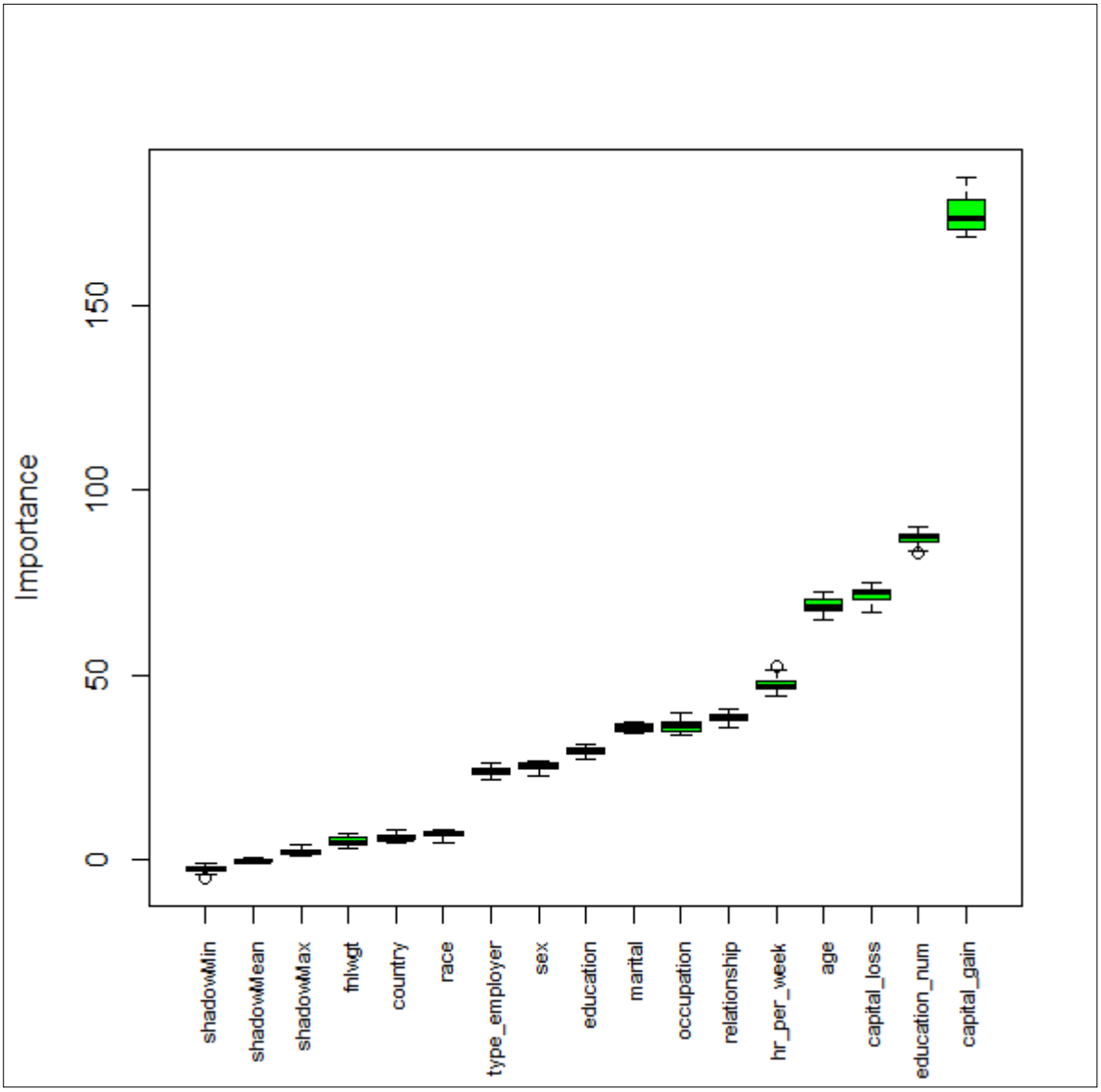
3. Feature Engineering

Select Data

In order to select a classifier for data preparation, a ROC curve was generated for multiple classifiers including Naïve Bayes, Rule Induction, kNN and Decision Tree with Naïve Bayes achieving the highest AUC on the unmodified training dataset.

These classifiers had been selected as suitable candidates as the dataset has a mix of nominal and numeric attributes with a binomial label.

Feature importance (R Package Boruta)



```
                meanImp  medianImp     minImp     maxImp  normHits  decision
age           68.957534  68.669511  65.257354  72.686096 1.0000000 Confirmed
type_employer 23.966407  23.955789  21.972548  26.184487 1.0000000 Confirmed
fnlwgt         5.110770   4.755538   3.075415   7.464417 0.8888889 Confirmed
education     29.661506  29.588748  27.566921  31.343032 1.0000000 Confirmed
education_num 87.074915  87.380765  82.959497  90.283987 1.0000000 Confirmed
marital       35.914397  35.882811  34.591738  37.537794 1.0000000 Confirmed
occupation    36.467786  36.561258  33.869914  39.877833 1.0000000 Confirmed
relationship  38.571582  38.354398  35.958424  41.061614 1.0000000 Confirmed
race           7.107547   7.354943   4.865782   8.321392 1.0000000 Confirmed
sex           25.500381  25.673310  22.860007  26.916362 1.0000000 Confirmed
capital_gain 174.617286 173.273459 168.218469 184.362607 1.0000000 Confirmed
capital_loss  71.887486  72.363071  67.118717  74.804402 1.0000000 Confirmed
hr_per_week   47.722864  46.970891  44.283630  52.225129 1.0000000 Confirmed
country        6.259687   5.827983   4.746966   8.248269 1.0000000 Confirmed
```

```
[1] "age"           "type_employer" "fnlwgt"        "education"
 [5] "education_num" "marital"       "occupation"    "relationship"
 [9] "race"          "sex"           "capital_gain"  "capital_loss"
[13] "hr_per_week"   "country"
```

Unmodified training dataset: Initial ROC curves on training dataset

To investigate the applicability of feature reduction to the full training dataset forward selection was applied to a Naïve Bayes classifier (Table 6 A). The results of this test indicate that although the >50K class precision was improved slightly over Naïve Bayes (without forward selection) the overall performance was degraded slightly (Table 6 B). The application of forward selection did however provide useful information as the fnlwgt (survey weight) and country attributes had been dropped without significantly impacting the performance of the Naïve Bayes classifier confirming the earlier assertion during data exploration that these attributes may not be significantly predictive.

Singular value decomposition (SVD) and principle component analysis (PCA) were then applied to the dataset in two ways. Initially SVD and PCA (with an optimal quantity of three dimensions/ components) were applied to the numerical attributes only which were joined to the nominal attributes and passed to a Naïve Bayes classifier with both achieving an eighty two percent accuracy (Table 6 C/E). SVD and PCA (with an optimal quantity of six dimensions/components)

were then applied to all attributes, where the nominal attributes had been mapped to real values, and both approaches now achieved an accuracy of seventy nine percent (Table 6 D/F). These results indicate that the above approaches to feature reduction on the unmodified training dataset do not appear to be useful in improving classifier accuracy.

Naïve Bayes with forward selection & PCA on training dataset.

Construct Data

When the marital status and relationship attribute values were modified slightly to reduce the number of categories (Table 7) it was observed that the information in the marital status attribute could be inferred to some degree from the 'unmarried' category in the relationship attribute suggesting that marital status might be a candidate for removal.

Discretisation was applied to the numerical attributes to determine if the performance of the Naïve Bayes classifier could be improved. Initially entropy binning was investigated on the training dataset but a performance of eighty two percent was poorer than that achieved without binning. However the output (Table 8) from this exercise did provide a useful starting point in suggesting bin boundaries for the following discretisation work.
16

various bin boundary combinations using Rapidminer's Optimise Parameter operator an optimal bin quantity for the hrs_per_week (worked) attribute was found to be twenty with the most highly populated bin (by a factor of ten) having bin boundaries of thirty five and forty with a mean of thirty nine. It was found that using a two bin approach with a bin boundary of this mean value of thirty nine proved to be equally effective. (Table 9).

Using the binning operator in rapidminer the optimal bin quantity for the age attribute was found to be eleven (Table 9) and based on the generated bin boundaries and work with R to determine the data distribution twenty bins were selected at boundaries ranging from twenty to sixty six (specifically 20, 25, 31, 36, 40, 46, 51, 56, 60, 66 and over 66) which slightly improved the classifier's performance. The more variate capital gain and loss attributes were found to have optimal binning quantities of five hundred and fifteen hundred respectively as increasing the bin quantity beyond this points did not improve the model's performance significantly (Table 9).
17ith Naïve Bayes it was found that removing the country (which has the same value in ninety percent of cases), education number (which is correlated with education level), survey weight (which is highly variate) and marital status (which appears to contain similar information to the relationship attribute as outlined earlier) attributes did not affect the model's performance. Therefore these four attributes will be considered for removal during

future modelling. The results of these initial data transformations are summarised below and will be referred to in this report as data transformations type A (Figure 9).

4. Modelling

Select modelling technique

As stated above as the dataset has mixed numerical and nominal attributes with a binomial class label a Decision Tree, kNN, Naïve Bayes and Rule Induction classifier had been selected for initial modelling.

Generate Test Design

During modelling the training dataset will initially be used to evaluate each classifier's performance relative to that achieved by Naïve Bayes on the unmodified training dataset (Table 6 A) and the optimal classifiers will then be evaluated on the test dataset in terms of overall performance and ability to support the primary business objective of maximising the return on investment.

Build and assess the model

As described in the data preparation section above applying forward selection on the unmodified training dataset with Naïve Bayes had not been very successful (Table 6) it was decided during modelling to revisit this idea but this time the data transformations (Type A) above were applied ahead of forward selection. This approach proved to be quite successful with a Naïve Bayes performance improvement of three percent (Table 10). Examination of the model's example-set showed that forward selection on the transformed data had also additionally removed the hrs_per_week, age, occupation and gender attributes which will be referred to as data transformations type B in this report (Figure 10).

Unmodified training dataset: Initial ROC curves on training dataset					
Model	**Accuracy**	**Prec**	**Recall**	**KAPPA**	**AUC**
GLM					
Random Forest					
Neural Network					
CART					

Unmodified training dataset: Initial ROC curves on training dataset					
Model	**Accuracy**	**Prec**	**Recall**	**KAPPA**	**AUC**
GLM					
Random Forest					
Neural Network					
CART					

Model	**Accuracy**	**Prec**	**Recall**	**KAPPA**	**AUC**
XGB (WITHOUT MISSING DATA)					
XGB (WITH INPUTATION)					
XGB (WITH HYPERPARAMETER TUNING)					

5. Evaluation

During the CRISP-DM data understanding phase of the project some of the attributes were found to be predominantly single valued such as the country (with 'United-States') and capital gain and loss (with 0) in over ninety percent of instances.

The survey weight attribute had the opposite issue where there was a large number of unique values. Later work with forward selection on Naïve Bayes confirmed that the country and survey weight attributes could be removed successfully without impacting classifier performance. The numeric education number and nominal education level attributes were found to be fully correlated (with education number being removed later during data preparation), and a correlation was also discovered between the marital status and relationship attributes after some basic category aggregation had been carried out and also between categories in the occupation attribute and the >50K class label.

The overall data quality was quite good with a low occurrence of conflicting attribute values and with over thirty thousand clean instances in the training dataset and fifteen thousand in the training dataset (with both having low levels of data duplication) there was a sufficient quantity of clean variant data in both class labels to successfully create (with training data) and evaluate (with test data) the various classifiers.

Some outliers were detected in the data but this was not found to seriously impact classifier performance as the percentage of affected instances was low in general with the hours_per_week (worked) attribute having slightly elevated levels but as binning was found to be beneficial on this attribute the impact of outliers was reduced.

The percentage of records with missing values was quite low with country at two percent and employment class and occupation at four percent which did not affect the classification work except in the case of kNN where the affected instances were successfully removed.

Some biases were detected in the data such as female (gender attribute) instances generally having a lower set of values than male instances but incidences of this type were not found to be significant.

The general observations made during data exploration were confirmed when an initial rule induction classifier was applied to the training data and again when the NBTree output was analysed. When the models which were deemed to be appropriate for this dataset (with mixed numerical and nominal attributes and a binomial label) were applied to the unmodified training data the Naïve Bayes classifier had the best performance.

During initial data preparation it was found that the optimal data transformation for Naïve Bayes involved discretising the hrs_per_week (worked), age, capital gain and capital loss attributes and removing the country, education number, survey weight and marital status attributes (which had been noted as possible candidates for removal earlier). This useful data transformation was tagged as type A in this report. When the type A data transformation was applied ahead of forward selection on a Naïve Bayes classifier a further performance improvement was achieved by additionally removing the hrs_per_week, age, occupation and gender attributes which was tagged as type B data transformation.

The type A data transformations were then applied to the hybrid DTNB and NBTree classifiers on the training dataset with improved performance on both and a slight modification to the type A (where the discretisation on the capital gain and loss attributes was removed as were instances with missing values) proved optimal for the kNN classifier which was tagged as type C data transforms.

The modelling work on the training dataset appears to indicate that there is a classifier accuracy limit of just below eighty seven percent which is supported by other work such as the Naïve Bayes performance given at eighty three to four percent with discretisation (Kaya, 2008) and also at eighty four percent (Kohavi, 1996) and NBTree results posted by UCI (UCI Archive, 2011) indicating a performance of eighty five percent. It therefore appears that the proposed models and associated data transforms outlined in this report are close to optimal.

During the modelling phase the training dataset had been used to evaluate the relative performance of each of the selected classifiers and these models were then evaluated on the test dataset on the level of accuracy and level of support (based on lift) for the primary goal of maximising return on investment. NBTree had the highest accuracy at 85.93% and was also the most profitable (on the current test dataset) when the first three (of ten) lift bins were used. The kNN classifier had the highest return on investment on all lift bins. The 'No Model' approach resulted in losses for all bins.

Based on these results it would seem that Naïve Bayes (with type B data transformations) offers good performance with minimum runtime and may be a good choice for the initial validation of new datasets. The maximum profit levels were generated by the NBTree classifier (with type A data transformations) as it reached more prospects than the other classifiers with a slightly longer runtime then Naïve Bayes. The maximum return on investment was generated by the kNN classifier (with type C data transformations) with a k value of 6 but this carried a runtime overhead which was five times longer than NBTree.

As the goals of identifying useful classifiers to support the business objective of maximising return on investment and the provision of descriptions of the attributes which were deemed to be significant have been achieved it is believed that the results outlined above satisfactorily support the stated business objectives and should form a good foundation for future classification work on similar datasets.

6. Appendix A

As an example of the calculations used to generate the return on investment and profitability charts we will work through the figures for the first bin quantities found with the xgboost model. The calculations are based on an initial setup cost of eighteen thousand dollars, with an offer cost of one hundred and twenty five dollars and an acceptance revenue of five hundred dollars:

XGBOOST model 1:

Number of prospects in bin = 1614
Predicted number of salaries > 50K = 1378

Predicted number of acceptances = 1378 * 0.90 = 1033
(*seventy five percent of individuals with salaries exceeding fifty thousand dollars are expected to accept the current offer strategy*)

Total cost = setup cost + total cost of offers sent
=18000 + (1614 * 125)
=219,500

Total revenue = number of acceptances * revenue per acceptance
=1033 * 500
=516,500

Profit = total revenue - total cost
= 516500 - 219500
= **297,000**

ROI = (profit / total cost) * 100
= (297000/219500) * 100
= **135**

7. References

UCI Archive, 2011, *http://archive.ics.uci.edu/ml/machine-learning-databases/adult/ adult.names*
30
https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

http://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html

https://www.r-bloggers.com/an-introduction-to- xgboost-r-package/